Document Classification with Reuter50 News Text: Update 2

Muzhe Zeng
mzeng6@wisc.edu

1 Dataset

Text classification task with 50 balanced categories. See this for details.

2 Current Progress

- Refine the data processing. Removing stopwords using nltk package.
- Add tentative experiments for the first two machine learning methods: multi-class logistic regression (LR) [1] and support vector machine (SVM) [2]. LR obtains 85% accuracy in the test set, while SVM only achieves 77% accuracy.
- Extract low rank feature from the document-term matrix using `vsp` [3]. The accuracy is slightly lower but the speed is much faster.

3 Future Plan

Implement deep learning (DL) to compare the performance. For SVM and DL, there many hyper-parameters to tune using cross-validation. Thus there improving space for SVM comparing to LR. I'll finish all these tuning and finalize the comparison before the deadline.

4 Github Link

`https://github.com/MuzheZeng/Document-Classification-with-vsp`

References

[1] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[2] Cortes, Corinna, and Vladimir Vapnik. "Support vector machine." Machine learning 20, no. 3 (1995): 273-297.

[3] Rohe, Karl, and Muzhe Zeng. "Vintage Factor Analysis with Varimax Performs Statistical Inference." arXiv preprint arXiv:2004.05387 (2020).